Étude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle

Clémentine Adam Cécile Fabre Ludovic Tanguy
CLLE-ERSS : CNRS & Université de Toulouse
5, Allées Machado, 31058 Toulouse Cedex 9
adam@univ-tlse2.fr, cfabre@univ-tlse2.fr, tanguy@univ-tlse2.fr

RÉSUMÉ

Dans cet article, nous confrontons une base distributionnelle construite à partir d'un corpus d'articles de revues de sciences humaines à des substitutions observées dans les journaux de requêtes du moteur interrogeant ce même corpus ; le recouvrement entre les deux types de données est important (59%). Ces résultats contribuent à deux pistes de recherche : d'une part nous montrons l'adéquation de la sémantique distributionnelle pour appréhender une large palette de relations sémantiques en jeu dans les reformulations de requêtes ; d'autre part, nous introduisons des données pouvant être exploitées pour l'évaluation de ressources distributionnelles de manière bien plus satisfaisante que par la comparaison avec des *gold standards* tels que des dictionnaires de synonymes.

ABSTRACT

Studying semantic relations in query reformulation under the scope of distributional analysis

In this paper, we compare a distributional resource built from a corpus of humanities and social sciences academic papers to substitutions recorded in user query logs covering the same corpus. We observed a good overlap between the two datasets (59%). These results show that distributional semantics is a fitting tool to analyze the wide variety of semantic relations involved in query reformulation. Moreover, the method that we introduce may be used for distributional resources evaluation, and is better fitted to this task than comparison with gold standards.

MOTS-CLÉS : sémantique distributionnelle, Recherche d'information, reformulation de requêtes.

KEYWORDS: distributional semantics, information retrieval, query reformulation.

1 Introduction

Les ressources lexicales construites par des méthodes distributionnelles ont été confrontées de différentes manières à l'expertise humaine pour chercher à cerner la nature des relations qui sont détectées par ce biais. La comparaison à d'autres ressources lexicales servant de *gold standard* – dictionnaires de synonymes et thésaurus – (Weeds, 2003; Anguiano *et al.*, 2011) a montré que les liens de proximité sémantique contruits de manière distributionnelle ne recouvrent que très partiellement les relations décrites dans les ressources lexicales existantes. Ainsi, (Morlane-Hondère et Fabre, 2012) montrent qu'on retrouve dans les couples de voisins distributionnels calculés à partir d'un corpus issu de Wikipédia 30 à 40% des couples contenus

dans un dictionnaire de synonymes, mais que ces synonymes ne constituent qu'une infime proportion de l'ensemble des liens de voisinage distributionnel. Ces proportions sont comparables lorsque la confrontation porte sur le réseau lexical français Jeux de mots (Lafourcade et Joubert, 2010), qu'il s'agisse des relations d'antonymie, d'hyperonymie, ou de méronymie. Par conséquent, l'utilisation de telles ressources n'éclaire qu'une toute petite partie seulement des relations sémantiques calculées. Deux raisons expliquent cette intersection limitée : le fait que d'autres relations de sens plus lâches, souvent qualifiées de "non classiques" (Morris et Hirst, 2004) et qui n'entrent ordinairement pas dans le périmètre des ressources lexicales existantes, sont détectées par la proximité distributionnelle (cohyponymie, mots de même domaine, etc.) ; le fait également que les relations acquises à partir de corpus peuvent être fortement contextualisées et par conséquent ne pas être recensées dans des lexiques. Une seconde approche consiste à confronter les résultats du calcul distributionnel au jugement humain à travers différentes tâches impliquant un jugement de similarité sémantique : détection de synonymes, *priming*, catégorisation, analogie (Padó et Lapata, 2007; Baroni et Lenci, 2010). Ces expériences montrent que la proximité distributionnelle offre une bonne approximation du jugement de similarité, et qu'une palette importante de relations sont détectées par la méthode distributionnelle.

Dans le travail présenté ici, notre objectif est de confronter une ressource distributionnelle à des données obtenues dans le contexte d'une tâche de recherche d'information. L'utilisation de ressources distributionnelles en RI a donné lieu à différentes expériences visant particulièrement l'expansion automatique de requêtes, depuis (Grefenstette, 1994) et, plus récemment, (van der Plas et Tiedemann, 2008; Picton *et al.*, 2008). Ces deux études, qui exploitent les données des campagnes d'évaluation CLEF, font état de résultats globalement décevants, avec des augmentations de performance non significatives. Elles montrent également par des analyses plus locales que les effets de l'expansion sont très contrastés, selon le type de requête, la catégorie de mots expansés et la nature du lien sémantique : les deux études mettent au jour des faiblesses spécifiques concernant l'expansion des adjectifs et des noms propres, liées en particulier à la génération de liens de cohyponymie qui introduisent du bruit (ex : lien de *russe* vers *allemand* ou *chinois*). L'exploitation de liens sémantiques non classiques (ex : *grève*/*réclamer*, *clonage*/*éthique*), et plus spécifiquement de liens entre des mots de même domaine (comme *bombe*, *guerre*) améliore par contre nettement le traitement de certaines requêtes. Une utilisation en aveugle de ressources distributionnelle s'avère donc inefficace. Une meilleure connaissance des types de requêtes et des relations de sens mises en jeu dans ce type de tâche s'avère nécessaire pour permettre une utilisation plus adaptée d'informations sémantiques en RI.

L'étude que nous présentons ici se situe en amont de ce type de travaux et se fonde sur des données écologiques : il s'agit de journaux de requêtes produites par les utilisateurs d'un moteur de recherche. L'objet de l'article est d'étudier le recouvrement entre les liens sémantiques exploités dans les reformulations, et ceux qui sont calculés sur la base du critère de proximité distributionnelle. Nous présentons d'abord les caractéristiques générales de ces données (section 2), avant de présenter l'échantillon de requêtes sur lequel se base notre étude : issu du moteur de recherche d'`openedition.org`, il est composé de couples de requêtes correspondant à des reformulations fondées sur la substitution d'un mot (section 3.1.). Nous présentons ensuite la base distributionnelle que nous avons construite à partir de la base de textes interrogée par les utilisateurs (section 3.2.). Dans la dernière section, nous présentons les résultats de cette comparaison, qui montrent un recouvrement important entre les deux sources de données (mots substitués / voisins distributionnels), et nous analysons certaines sources de décalage permettant de mieux comprendre les différences entre ces deux types de proximité sémantique.

2 Les reformulations de requêtes

L'étude des journaux des requêtes soumises à un moteur de recherche (ou *query logs*) a, dès le début, suscité un grand intérêt pour la communauté de la recherche d'information, que ce soit pour mieux comprendre le comportement des utilisateurs ou pour modéliser celui-ci afin d'améliorer la réponse d'un moteur de recherche (Jansen *et al.*, 2000).

Parmi les différents phénomènes abordés dans ces études, la notion de *reformulation* est l'une des plus riches en information sur la nature des interactions entre un utilisateur et un système de recherche (Jansen *et al.*, 2009). Une reformulation est simplement la soumission au moteur de recherche d'une seconde requête, différente de la précédente, mais qui cherche à remplir le même besoin d'information. Le temps passé et les actions de l'utilisateur entre les deux requêtes peuvent varier, mais on a pu dégager empiriquement les principales modifications formelles que subissent les textes des deux requêtes. En nous inspirant de l'état de l'art établi par (Huang et Efthimiadis, 2009), nous les présentons dans la table 1 avec des exemples extraits de notre propre collection de données.

Modification	Exemples
Correction orthographique	`praglatismes` → `pragmatisme`
Modification de l'espacement	`église catholiquefinancement` → `église catholique financement`
Réordonnancement	`monde rural france survivance` → `survivance monde rural france`
Ajout de mot(s)	`algerie` → `algerie femmes`
Suppression de mot(s)	`hyperactivité à l'ecole` → `hyperactivité`
Remplacement de mot(s)	`stéréotypes sexistes` → `stéréotypes machistes`

TABLE 1 – Différents types de schémas de reformulation

D'un point de vue plus profond, les reformulations qui modifient de façon non triviale l'expression de la requête peuvent être classées suivant les 4 types suivants, qui pour certains se traduisent par plusieurs mécanismes de surface : *généralisation*, *spécification*, *reformulation* et *mouvement parallèle*. Des exemples de telles opérations sont présentés dans la table 2. La généralisation correspond clairement à un élargissement du champ de recherche visant à réduire le silence, et la spécification à son rétrécissement pour réduire le bruit. Ces deux opérations peuvent se traduire chacune par deux opérations formelles (remplacement ou ajout/suppression de mots). La reformulation correspond à une simple paraphrase, alors que le mouvement parallèle est au contraire une modification importante de la requête, en considérant une alternative [1].

Comme on peut le voir dans ces quelques exemples, ces reformulations sont a priori le lieu d'expression de relations sémantiques très variées entre les requêtes (ou entre les termes de celles-ci). Plus particulièrement, on peut voir dans la table 2 que les opérations de substitution d'un mot par un autre peuvent se faire en suivant (au moins) quatre relations sémantiques : hyponymie, hyperonymie, synonymie et co-hyponymie. Dans le cas de cette substitution, le fait qu'une partie de la requête reste inchangée est un indicateur fiable de la continuité sémantique entre les deux requêtes, même si la cible de la requête a pu être sensiblement modifiée.

Les requêtes sur lesquelles nous basons cette étude, et dont sont extraits les exemples présen-

1. Les exemples canoniques du mouvement parallèle sont, dans le cas des requêtes soumises aux moteurs Web généralistes, le remplacement d'une marque ou d'un type de produit par un autre lors d'une recherche à visée transactionnelle, comme dans `nikon camera` → `canon camera` ou `billet train paris-toulouse` → `billet avion paris-toulouse`.

Généralisation	Suppression de mots	`inégalités sociales de santé` → `inégalités de santé`
	Remplacement par un hypéronyme	`sociologie arts martiaux` → `sociologie du sport`
Spécification	Ajout de mots	`faillite` → `théories de la faillite`
	Remplacement par un hyponyme	`activités dans la montagne` → `escalades dans la montagne`
Reformulation	Remplacement par un synonyme	`territoire voiture` → `territoire automobile`
Mouvement parallèle	Remplacement par un co-hyponyme	`siège romain` → `siège gaulois`

TABLE 2 – Quatre types d'opération effectuée lors de la reformulation de requêtes

tés ci-dessus, sont celles qui ont été soumises au moteur de recherche du site OpenEdition (`www.openedition.org`) qui regroupe un grand nombre de ressources documentaires dans le domaine des sciences humaines et sociales. Nous disposons, grâce à la collaboration du CLEO (Centre pour L'édition Electronique Ouverte) de l'ensemble des requêtes soumises au moteur, ainsi que des adresses anonymisées qui nous permettent de suivre le comportement d'un utilisateur au cours d'une session (documents consultés, temps passé, etc.), voir (Leva, 2013) pour plus de détails.

3 Données et méthodologie

Dans cette section, nous décrivons les données que nous avons extraites et confrontées dans le cadre de cette étude : un ensemble de paires de requêtes présentant des remplacements lexicaux, extraites des journaux de requêtes d'OpenEdition (sous-section 3.1) et une base distributionnelle construite à partir de la collection de documents interrogée par le moteur de recherche d'OpenEdition (sous-section 3.2). Nous présentons ensuite la méthodologie de croisement de ces données (sous-section 3.3).

3.1 Les substitutions dans les journaux de requêtes

Pour l'extraction de paires de requêtes comprenant une substitution, nous avons utilisé un extrait des logs de recherche d'OpenEdition comprenant 194 363 requêtes réparties en 57 813 sessions de recherche de plus de deux requêtes (une session est ici définie grossièrement comme une séquence de requêtes soumises par le même utilisateur dans une même journée).

À partir de ces données, nous avons extrait toutes les paires de requêtes R_a, R_b telles que :
– R_a et R_b appartiennent à la même session et sont consécutives ;
– R_a et R_b comprennent un même nombre n de mots avec $n \geq 2$;
– R_a et R_b comprennent $n-1$ mots communs.
Nous nous basons donc sur une vision limitée de la substitution, qui ne porte ici que sur un seul mot dans des requêtes de plusieurs mots. Ces critères ont permis d'extraire un ensemble de 11321 paires de requêtes, ce qui représente plus de 8% des paires de requêtes du log analysé. Cela montre que le type de reformulations sur lequel nous avons choisi de travailler, bien que très circonscrit, n'est pas marginal dans les stratégies de recherche des utilisateurs.

Nous fournissons dans la suite de cette sous-section quelques éléments de description des paires de requêtes extraites, dont des exemples sont répertoriés dans le tableau 3.

a	relations entre espace pouvoir et **société** relations entre espace pouvoir et **identité**
b	la gouvernance **internationalle** la gouvernance **international**
c	québec intégration **immigration** travail intégration **immigrants** travail québec
d	**securité** des sites de rencontres **dangers** des sites de rencontres
e	traduction **inginierie** traduction **technique**
f	**performance** système d'information **évaluer** système d'information
g	anthropologie des **iécoles** anthropologie des **institutions**
h	représentation **culturelle** de la douleur représentation **sociale** de la douleur
i	desavantages de l apprentissage d une **deuxieme** langue desavantages de l apprentissage d une langue **etrangere**
j	lieux de pouvoir paris élysée **louvre** lieux de pouvoir paris élysée **matignon**
k	cheminées **hydroterhmales** cheminées **hydrothermales**
l	societé **et** transition societé **en** transition
m	ville **de** paris **imaginaire** ville paris
n	**encourir** une condamnation **prononcer** une condamnation
o	**emploi** rural **travail** rural
p	climat **mousson** climat **tropical**
q	rire **thérapeutique** rire **guérir**
r	musique et **apprentissage** **musique** et **mémorisation** **rythmique** et mémorisation
s	stéréotypes **genrés** stéréotypes **sexistes** stéréotypes **machistes**

TABLE 3 – Exemples de paires de requêtes extraites

Positions des mots substitués Les paires de requêtes extraites comprennent entre 2 et 20 mots (moyenne : 2,95 mots). Lors de l'extraction, aucune incidence n'a été donnée à l'ordre des mots dans chaque requête. Il apparaît toutefois que celui-ci est généralement préservé entre les deux requêtes (peu de réordonnancements). La figure 1 résume les schémas de substitution observés : dans 92% des cas, le mot supprimé et le mot ajouté se trouvent à la même position, majoritairement à la fin de la requête (52%). Les changements de position (*cf.* tab. 3 ex. c et i) sont rares sans être marginaux : 8% des paires de requêtes sont concernées.

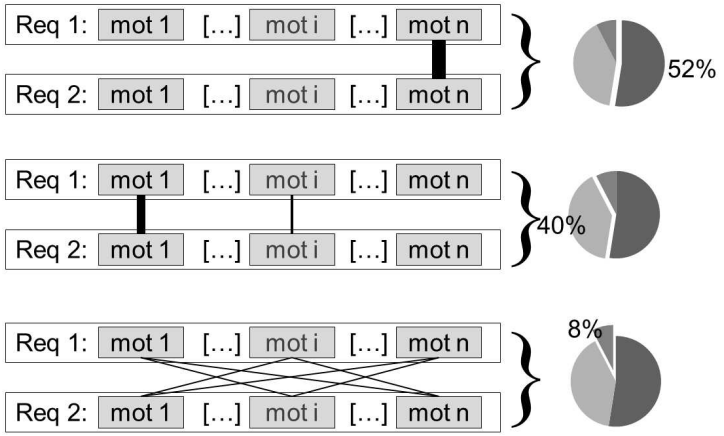

FIGURE 1 – Positions des mots substitués

Mots vides, corrections orthographiques et fautes non corrigées Certaines substitutions portent sur des mots vides (prépositions, déterminants, ...) : substitués entre eux (tab. 3 ex. l) ou avec un mot plein (ex. m) – dans ce dernier cas, il s'agit alors plutôt d'un ajout que d'un remplacement (le mot *imaginaire* est ajouté à la requête plus que substitué au mot *de*, qui n'avait pas d'impact sur la recherche). Nous avons repéré ces mots-vides à l'aide d'une *stop-list* ; 563 paires de requêtes sont concernées (moins de 5% de l'ensemble). Les corrections orthographiques, qui ne nous intéressent pas dans le cadre de cette étude, représentent quant à elles une très large partie des données extraites. Afin d'évaluer leur importance, nous avons appliqué une distance d'édition (ou distance de Levenstein) sur les paires de mots substitués, après exclusion des mots vides (les paires telles que `et` → `en` ne sont donc pas considérées comme des corrections orthographiques). Avec cette méthode, nous avons repéré 5155 corrections orthographiques (46% des données). Mais toutes les fautes d'orthographe ne donnent pas lieu à une correction, comme on peut le voir avec les substitutions `inginierie` → `technique` (ex. e) et `iécole` → `institution` (ex. g) ; ces paires exhibent un lien sémantique, mais sont de nature à faire buter les traitements. Pour repérer les mots incorrects, nous avons fait appel à deux lexiques de formes fléchies : Morphalou [2] et GLÀFF (Sajous *et al.*, 2013), ainsi qu'à une liste de noms propres extraite de Wikipédia. 1727 reformulations (15% des données) impliquent des mots qui ne sont recensés par aucune de ces trois ressources.

Catégories morpho-syntaxiques concernées À partir des données filtrées et lemmatisées par l'appel aux lexiques, nous avons étudié la représentation des catégories morpho-syntaxiques dans les reformulations de requêtes. La figure 2 résume les proportions observées. Les noms dominent largement, suivis des adjectifs, des noms propres et des verbes ; un très petit nombre d'adverbes a également été relevé (ex. `parler franchement` → `parler franc`). Les remplacements au sein d'une même catégorie, par exemple Nom→Nom (tab. 3 ex. a, c, d, etc.) sont logiquement majoritaires, Adjectif→Adjectif (ex. h, i). Toutefois, les substitutions intercatégorielles sont également très fréquentes : le tableau 3 donne à voir des exemples de substitutions Nom→Adj (ex. e et p), Nom→Verbe (ex. f) ou encore Adjectif→Verbe (ex. q).

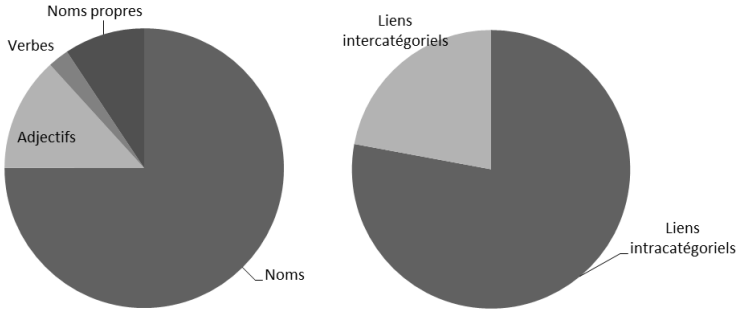

FIGURE 2 – Catégories morpho-syntaxiques des mots substitués

Variété des liens sémantiques Les reformulations qui nous intéressent plus particulièrement dans le cadre de cette étude sont celles qui présentent un lien sémantique entre les deux mots

2. développé par le CNRTL (ATILF) : `http://www.cnrtl.fr/lexiques/morphalou/`

substitués. Il s'agit donc, *a priori*, de toutes celles qui ne sont pas des corrections orthographiques ou des remplacements de mots vides, puisque nous faisons l'hypothèse que la partie inchangée lors de la reformulation est garante de la continuité sémantique entre les deux requêtes (*cf.* section 2). Les données présentées dans le tableau 3 donnent à voir une grande variété de relations sémantiques, allant au delà de ce qui est classiquement envisagé dans l'étude des reformulations de requêtes (hypo/hypéronymie, co-hyponymie, synonymie, *cf.* tableau 2). Outre les liens intercatégoriels déjà mentionnés (*performance*/*évaluer*, *mousson*/*tropical*, *thérapeutique*/*guérir*, etc.), on peut relever un cas d'antonymie (*sécurité*/*dangers*, ex. d) ainsi que des liens sémantiques plus lâches (par exemple *société*/*identité*, ex. a). Ces observations encouragent à aborder les reformulations de requêtes par le biais de la sémantique distributionnelle. Nous présentons dans ce qui suit la base distributionnelle utilisée et son croisement avec les paires de requêtes extraites.

3.2 La base distributionnelle utilisée : les Voisins de Revues.org

La base de voisins distributionnels utilisée dans cette étude a été construite à partir de documents francophones provenant de la collection interrogée par le moteur de recherche d'OpenEdition. Plus particulièrement, il s'agit d'articles de sciences humaines issus de revues en ligne éditées par Revues.org. Le corpus constitué fait environ 253 millions de mots. Il a été soumis à une chaîne de traitement impliquant :

– une analyse syntaxique (analyseur Talismane, (Urieli et Tanguy, 2013)) ;
– une analyse distributionnelle (module Upéry, (Bourigault, 2002)).

L'analyse distributionnelle effectuée s'appuie donc sur des contextes syntaxiques : un mot peut constituer un contexte pour un autre mot s'il est rattaché à lui par une relation de dépendance syntaxique (objet, sujet, modifieur) ; ces relations de dépendance peuvent être labellisées par une préposition (« à », « de », « sur », etc.). Deux types de rapprochements sont opérés :

– des rapprochements de prédicats, c'est-à-dire de mots gouverneurs de la relation syntaxique, auxquels cette dernière est accolée ; par exemple, *évaluer_obj* et *mesurer_obj* sont voisins car ces deux prédicats partagent des contextes tels que *impact*, *dangerosité*, etc. ;
– des rapprochements d'arguments, c'est-à-dire de dépendants syntaxiques ; par exemple *social* et *culturel* sont voisins car ils partagent des contextes tels que *tâche_de*, *favoriser_obj*, etc.

La similarité des distributions (c'est-à-dire des ensembles de contextes d'apparition) de deux mots est évaluée par le score de Lin (Lin, 1998). Dans le cadre de ce travail, nous considérons que deux mots constituent une paire de voisins lorsque leur score de Lin est supérieur à un seuil fixé à 0,1.

3.3 Croisement des paires de requêtes et des voisins

Afin de comparer les relations sémantiques mises en jeu dans les reformulations de requêtes et celles calculées par l'analyse distributionnelle du corpus, nous nous sommes appuyés sur deux jeux de données.

Pour le premier, nous avons exclu les corrections orthographiques, les reformulations portant sur des mots vides et les mots incorrects (*cf.* section 3.1). Nous avons également supprimé les doublons : par exemple, bien qu'apparaissant dans plusieurs paires de requêtes (*cf.* exemples (i) et (ii) ci-dessous), les substitutions *marketing*/*communication* et *enseignement*/*éducation* ne sont prises en compte qu'une seule fois (seule la première paire de requêtes est conservée).

(i) `marketing solidaire web 2 0` → `communication solidaire web 2 0`
`communication luxe` → `marketing luxe`
`marketing humanitaire` → `communication humanitaire`

(ii) `enseignement en milieux hospitaliers` → `éducation en milieux hospitaliers`
`enseignement des détenus mineurs` → `éducation des détenus mineurs`

Pour le second jeu, nous avons également supprimé les reformulations portant sur des mots absents de la base distributionnelle utilisée. Les éléments filtrés pour la constitution des deux jeux de données sont récapitulés dans le tableau 4.

Élements filtrés	Exemples		Nb
Mots-vides	`et` → `en`	(tab. 3 ex. i)	563
Corrections	`hydroterhmales` → `hydrothermales`	(tab. 3 ex. k)	5155
Mots inconnus	`iécoles` → `institutions`	(tab. 3 ex. g)	1727
Doublons	`marketing` → `communication` (x3)	(i)	399
	⇒ Premier jeu de données : 3481 paires de requêtes		

Mots absents de la base	`résistance quinine` → `résistance chloroquine`	709
	⇒ Second jeu de données : 2772 paires de requêtes	

TABLE 4 – Récapitulatif des éléments filtrés
Point de départ : 11321 paires de requêtes

4 Résultats et analyse

Le tableau 5 montre les résultats obtenus en croisant les deux jeux de données décrits dans la section 3.3 avec les voisins distributionnels.

	Paires de requêtes	Voisins	
Jeu 1	3481	1641	47%
Jeu 2	2772	1641	59%

TABLE 5 – Résultats : paires de mots substitués voisins

Ces résultats montrent tout d'abord l'intérêt de notre démarche pour les deux directions évoquées en introduction :
– le recouvrement entre les voisins distributionnels et le premier jeu de requêtes montre que la sémantique distributionnelle permet d'appréhender efficacement les relations de sens impliquées dans les reformulations de requêtes, ce qui laisse envisager d'utiliser le critère du voisinage distributionnel pour aider au typage d'autres schémas de reformulations ;
– les données choisies – qui exhibent une large gamme de relations sémantiques, sont adaptées au corpus et sont produites par des utilisateurs dans une situation concrète de recherche d'information – apparaissent comme une référence idéale pour l'évaluation (en rappel) d'une base distributionnelle. Le recouvrement observé avec le second jeu de données (59%) est supérieur à ce qui a pu être observé en confrontant des bases distributionnelles à des *gold standard* (Morlane-Hondère et Fabre, 2012).

Nous avons par ailleurs croisé la présence du lien dans la base distributionnelle et différents paramètres, et avons trouvé une liaison significative avec la position des mots substitués, qui ont davantage tendance à présenter un lien de voisinage lorsque leur position est identique d'une requête à l'autre que lorsqu'elle change (test du χ^2, $p < 10^{-7}$). Cet effet est toutefois marginal (ϕ=0.1) : 61% des mots substitués à la même position sont des voisins, contre 44% si la position change.

Notons que l'expérimentation menée est adaptée à une évaluation en rappel des voisins distributionnels, mais pas en précision, du fait du manque d'exhaustivité des données : le fait qu'une paire de mots ne soit pas exploitée dans les reformulations ne signifie pas qu'elle ne présente pas de relation sémantique pertinente. Il est toutefois important de mettre en perspective le rappel obtenu afin de déterminer s'il n'est pas « gonflé » par une surgénération de voisins (il est facile d'obtenir un rappel parfait, si la précision de la base distributionnelle est quasi-nulle). En effet, le nombre de voisins des mots impliqués dans les reformulations est important (2191 en moyenne par mot). Dans cet objectif, nous avons vérifié le rang auxquel apparaît, pour chaque mot donnant lieu à une substitution, la relation exploitée dans cette substitution parmi tous ses voisins (ordonnés suivant le score de Lin décroissant). Par exemple, dans les reformulations `web documentaire` → `web reportage` ou `les jeux didactiques` → `les jeux pedagogiques`, *reportage* et *pédagogique* sont premiers parmi les voisins de, respectivement, *documentaire* et *didactique*. Par contre, dans la reformulation `espace urbain` → `aménagement urbain`, *aménagement* n'est que le 1346e voisin d'*espace* (sur 3181 voisins). L'analyse des rangs montre que pour la moitié des reformulations, le mot ajouté apparaît dans les premiers 10% de voisins du mot remplacé, ce qui montre que le fort rappel observé repose surtout sur une proximité distributionnelle importante et ne peut être imputé à une simple surgénération de voisins.

Si une proportion de 59% peut être considérée comme importante pour des tâches de confrontation de l'analyse distributionnelle à des données réelles, elle est toutefois loin d'être pleinement satisfaisante eu égard aux conditions optimales de l'expérimentation menée. On peut alors se demander quels sont les éléments impliqués dans le « silence » observé, du point de vue des mécanismes mis en jeu dans le calcul distributionnel – des mots auraient-ils pu être rapprochés mais ne l'ont pas été ? – et de la nature des données exploitées – l'hypothèse de continuité sémantique entre les deux requêtes doit-elle être mitigée ? (ce que suggère la corrélation avec la position des mots substitués), certaines relations sont-elles moins captées ? Cela nous a amenés à examiner plus en détails deux aspects : (a) les contextes distributionnels des mots substitués, qu'ils soient voisins ou non voisins (sous-section 4.1) ; (b) la nature des relations sémantiques captées ou non par le voisinage distributionnel (sous-section 4.2).

4.1 Contextes distributionnels

Nous avons mis en place plusieurs procédures d'observation visant à expliciter les mécanismes distributionnels permettant de rapprocher les termes substitués dans les requêtes.

4.1.1 Identification des contextes à l'origine des relations de voisinage distributionnel

Dans un premier temps, pour en avoir un aperçu global, nous avons regroupé tous les contextes syntaxiques communs à deux mots substitués, sur l'ensemble des 2772 paires de requêtes pour

lesquelles les deux mots substitués sont présents dans la base distributionnelle. Au total, 24563 contextes syntaxiques différents ont été identifiés, que ceux-ci aient été suffisants ou non pour que les mots substitués soient déclarés comme voisins.

Nous avons calculé, pour chaque contexte ainsi identifié, le nombre de couples de mots substitués qu'il a permis de rapprocher. Le tableau 6 présente les contextes les plus fréquents. Nous y avons distingué les prédicats (à gauche) des arguments (à droite),

Prédicat	Nb. de couples
pouvoir_suj	1870
faire_suj	1634
devoir_suj	1471
concerner_obj	1421
se agir_de	1372
constituer_suj	1353
histoire_de	1258
devenir_suj	1247
forme_de	1207
sembler_suj	1206
apparaître_suj	1198
question_de	1197
mettre_obj	1187
aller_suj	1176
faire_de	1159
cas_de	1158
faire_obj	1139
rester_suj	1129
parler_de	1124
type_de	1068

Argument	Nb. de couples
tout	1441
tel	1352
nouveau	1338
autre	1272
politique	1267
seul	1144
français	1033
premier	1028
grand	1010
même	945
propre	935
page	931
différent	874
véritable	848
social	843
ancien	833
pays	819
nombreux	797
important	790
présent	745

TABLE 6 – Liste des contextes les plus fréquemment communs aux deux mots substitués

Comme on peut le voir, si certains contextes sont très génériques (sujets des verbes modaux, modifieurs *tout* ou *tel*), on y retrouve rapidement des situations spécifiques à un corpus de SHS, que ce soit par les thématiques abordées (*politique*, *histoire_de*, *français*, etc.) ou par les formes présentatives (*question_de*, *constituer_obj*, etc.).

La présence de l'argument *page* dans cette liste est un des artefacts liés à l'analyse syntaxique automatique du corpus : les syntagmes nominaux du type *page X* sont considérés comme étant des appositions, et donc rattachés à une très grande variété de noms. Par exemple, dans l'extrait *« Dans les temps modernes (page 36) [...] »*, *page* est à tort considéré comme un modifieur du nom *temps* par Talismane [3].

4.1.2 Recouvrement des contextes distributionnels avec le contexte des requêtes

Dans un second temps, nous avons regardé spécifiquement si les contextes distributionnels communs se retrouvaient également dans le texte commun aux deux requêtes consécutives.

3. Qui est un analyseur statistique qui, à sa décharge, n'avait jamais rencontré ce type de configuration dans son corpus d'apprentissage...

Autrement dit, nous avons identifié les cas où les mots inchangés de la paire de requêtes qui forment une reformulation sont, dans la base distributionnelle, des éléments contextuels qui permettent de rapprocher les deux mots substitués.

Par exemple, dans la paire de requêtes :
`ligne aérienne internet` → `compagnie aérienne internet`
l'adjectif *aérien* est un contexte commun à *ligne* et à *compagnie* dans la base : il y a donc recouvrement des contextes entre le corpus et la requête.

Il se trouve que 27% des paires de requêtes étudiées présentent cette caractéristique avec au moins un mot de la requête correspondant à un contexte distributionnel. On y trouve à la fois des contextes très génériques comme *histoire_de* dans le couple
`histoire de l'ethnologie` → `histoire de l'anthropologie`
ou *femme_mod* dans
`femme antique` → `femme romaine`,
mais il peut s'agir également de structures très figées comme *caritatif* dans
`oeuvres caritatives` → `organisations caritatives`
ou encore *pénal* dans
`droit pénal` → `procédure pénale`.

Dans certains cas (10% des reformulations pour lesquelles le voisinage distributionnel n'a pas été identifié), nous avons pu observer la présence d'un contexte commun très spécifique, qui n'a pas permis à lui seul de permettre le rapprochement des deux termes comme voisins. C'est le cas du prédicat *mariage_mod* pour la paire de requêtes
`mariage forcé` → `mariage arrangé`
ou encore de *culture_mod* pour
`culture vivrière` → `culture maraîchère`.

Ces premières observations nous confirment le rôle central du corpus sur lequel la base de voisins est construite, mais nous amènent aussi à requestionner le mode de calcul actuel (et notamment le seuillage) de la notion de voisinage.

4.2 Relations sémantiques

Nous avons cherché à caractériser les différences sur le plan des relations sémantiques entre les deux séries de résultats : les couples de mots substitués qui sont des voisins distributionnels et ceux qui n'en sont pas. Pour cela, nous avons annoté deux échantillons de 200 couples, en leur attribuant les étiquettes qui sont mentionnées dans le tableau 7.

Cette annotation ne peut fournir que des indications très sommaires. On sait en effet la difficulté à annoter des relations sémantiques hors contexte (Adam, 2012). Nous pouvons néanmoins dégager deux types d'observation.

Tout d'abord, ces chiffres donnent un aperçu de la nature des liens qui sont mobilisés dans les reformulations. Ils confirment la prépondérance des liens sémantiques non classiques. Les cinq premières relations listées dans le tableau concernent seulement 1/3 des mots substitués, l'hyperonymie (généralisation ou spécification) et la cohyponymie (mouvement parallèle) étant les relations dominantes. On voit en effet que les relations dominantes entre mots substitués sont plus difficiles à caractériser et témoignent d'un saut conceptuel plus important pour atteindre

Nature du lien sémantique	Exemples	V (%)	non V (%)	tous (%)
synonymie	emploi-travail, mensonger-trompeur	11	3	7
lien morphologique	urbanisme-urbanisation, sacralité-sacré	3,5	6	4,75
antonymie	exclusion-intégration, sécurité-danger	2	0,5	1,25
hyperonymie	mariage-alliance, drogue-cannabis	13,5	6,5	10
méronymie	Angola-Afrique, puce-mémoire	0,5	3	1,75
cohyponymie	guadeloupe-martinique, reportage-documentaire	9	7,5	8,25
champ lexical	enfant-école, politique-engagement	25	31	28
autre lien	urbanisation-attractivité, attention-motivation	28,5	22	21,75
absence de lien	médiation-mine, banque-antécédent	7	20,5	13,75

TABLE 7 – Relations sémantiques identifiées dans les résultats substitués

un mot du même champ lexical voire un lien associatif plus lâche. On peut s'étonner de trouver une proportion non négligeable de mots substitués sans lien sémantique évident. Cette situation correspond à plusieurs cas de figure illustrés ci-dessous. Les requêtes (i) et (ii) correspondent à un double mouvement d'effacement et d'ajout, plutôt qu'à une substitution (on constate d'ailleurs un changement de position entre le mot originel et le mot substitué). Les exemples (iii), (iv) et (v) montrent des cas de redénomination qui rendent difficiles la mise en correspondance des composants substitués (*américanisation* et *exception* présentent une relation de contraste dans ce seul contexte très particulier). Enfin, les requêtes (v) et (vi) illustrent des cas de modification importante dans la désignation de la cible de la requête.

(i) `Maurice Genty` → `Genty contoverses`
(ii) `extériorisation logistique` → `risques extériorisation`
(iii) `Web 2.0` → `ville 2.0`
(iv) `jeux olympiques 2008` → `jeux olympiques Pékin`
(v) `exception culturelle` → `américanisation culturelle`
(vi) `mimétisme alimentation` → `media alimentation`
(vii) `art engagé` → `art osé`

Ces premières observations confirment la nécessité d'identifier des liens sémantiques variés lorsqu'il s'agit de mettre en place des procédures visant à accompagner les procédures de reformulation. Ces chiffres montrent aussi que l'ensemble des relations sont captées, dans des proportions néanmoins très variables, par le critère de proximité distributionnelle.

On peut par ailleurs relever les différences les plus nettes entre les deux échantillons. Elles concernent la part de synonymes (22 des 28 synonymes ont été repérés par le voisinage distributionnel), et la proportion de couples sans relation sémantique : ces deux chiffres attestent du rôle de filtre du voisinage distributionnel, qui repère les liens sémantiques les plus étroits et écarte les mots sémantiquement éloignés. Ce filtre n'est pourtant pas totalement efficace. Le non repérage de certains synonymes s'explique en particulier par des fréquences très déséquilibrées entre les deux mots (ex : *litige*/*conflit*, *herméneutique*/*interprétation*). De même, le voisinage distributionnel contribue fortement à filtrer les couples sans lien sémantique identifiable mais il ne les exclut pas totalement. Ainsi, les mots *médiation* et *mine* sont rapprochés parce qu'ils peuvent tous deux figurer en complément de certains noms très généraux (*travail*, *champ*, *produit*, *service*, *développement*, *système*...). La plupart de ces couples sont eux-mêmes composés de mots très généraux, aux contextes par conséquent peu discriminants (*réseau*/*genre*, *science*/*gouvernance*, *démocratie*/*pensée*, *type*/*technique*, *monde*/*mouvement*...).

5 Conclusion et perspectives

Cette étude a montré que le critère distributionnel permet de détecter des types variés de proximité sémantique, qui sont mobilisés dans l'activité de reformulation que pratiquent des utilisateurs d'un moteur de recherche : les opérations de substitution sémantique qui sont utilisées dans les requêtes font appel à des relations de sens multiples, et majoritairement à des relations lâches qui sont bien détectées par le critère du voisinage distributionnel (cohyponymes, mots appartenant au même champ lexical, liens très contextualisés). L'analyse des résultats a permis par ailleurs de mettre au jour les faiblesses de l'analyse distributionnelle sur deux points principaux : la surgénération de voisins à partir de contextes partagés trop généraux ; la non détection de mots proches qui ne sont liés que par une série très limitée de contextes, voire un contexte unique (comme *forcé* et *arrangé* modifiant le seul mot *mariage*).

Si l'on se place dans une perspective de recherche d'information (avec un objectif à long terme d'assistance de l'utilisateur ou de détection de reformulation), on peut considérer que des ressources du type des bases distributionnelles sont effectivement adaptées. Toutefois, leur mode de calcul et de projection doit être spécifiquement affiné pour lutter contre le bruit que génère ce type de méthode. On peut également envisager d'exploiter les informations contextuelles des requêtes (ne serait-ce que l'ensemble des mots de celles-ci, ou l'historique des requêtes d'un utilisateur) pour combler les lacunes des calculs distributionnels.

Remerciements

Ce travail s'inscrit dans le cadre du projet ANR CAAS (*Contextual Analysis and Adaptative Search*), programme CONTINT (2010-2014), coordonné par Josiane Mothe (IRIT).

Merci à Marin Dacos, à Patrice Bellot et à toute l'équipe du CLEO pour nous avoir permis d'accéder à leurs logs de requêtes ainsi qu'à leur corpus.

Merci à Simon Leva (CLLE) et à Nicolas Faessel (IRIT) d'avoir filtré, croisé et dépouillé les différents logs d'accès.

Merci à Franck Sajous (CLLE-ERSS) pour son travail de longue haleine sur la construction des bases distributionnelles.

Références

Adam, C. (2012). *Voisinage lexical pour l'analyse du discours*. Thèse de doctorat, Université de Toulouse.

Anguiano, E., Denis, P. *et al.* (2011). Fredist : Automatic construction of distributional thesauri for french. *In Actes de TALN*, pages 119–124.

Baroni, M. et Lenci, A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN*, pages 75–84, Nancy.

GREFENSTETTE, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Dordrecht : Kluwer Academic Publishers.

HUANG, J. et EFTHIMIADIS, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *In Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86. ACM.

JANSEN, B., SPINK, A. et SARACEVIC, T. (2000). Real life, real users, and real needs : a study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227.

JANSEN, B. J., BOOTH, D. L. et SPINK, A. H. (2009). Patterns of query reformulation during web searching. *American Society for Information Science and Technology Journal*, 60(7):1358–1371.

LAFOURCADE, M. et JOUBERT, A. (2010). Construction de l'arbre des usages nommés d'un terme dans un réseau lexical évolutif. *Actes des Journées internationales d'Analyses statistiques des Données Textuelles (JADT'10)*.

LEVA, S. (2013). Les sessions de recherche comme contexte des requêtes. *In Actes de l'atelier sur la Contextualisation des Messages Courts, dans le cadre de EGC*, pages 1–13, Toulouse.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, Madison.

MORLANE-HONDÈRE, F. et FABRE, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *In Actes du Congrès Mondial de Linguistique Française*.

MORRIS, J. et HIRST, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT Workshop on Computational Lexical Semantics*, pages 46–51, Boston.

PADÓ, S. et LAPATA, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

PICTON, A., FABRE, C. et BOURIGAULT, D. (2008). Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation du voisinage distributionnel. *Revue Française de Linguistique Appliquée*, XIII(1):83–96.

SAJOUS, F., HATHOUT, N. et CALDERONE, B. (2013). Glàff, un gros lexique à tout faire du français. *In Actes de TALN*.

URIELI, A. et TANGUY, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur talismane. *In Actes de TALN*.

van der PLAS, L. et TIEDEMANN, J. (2008). Using lexico-semantic information for query expansion in passage retrieval for question answering. *In Coling 2008 : Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 50–57. Association for Computational Linguistics.

WEEDS, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. Thèse de doctorat, University of Sussex.